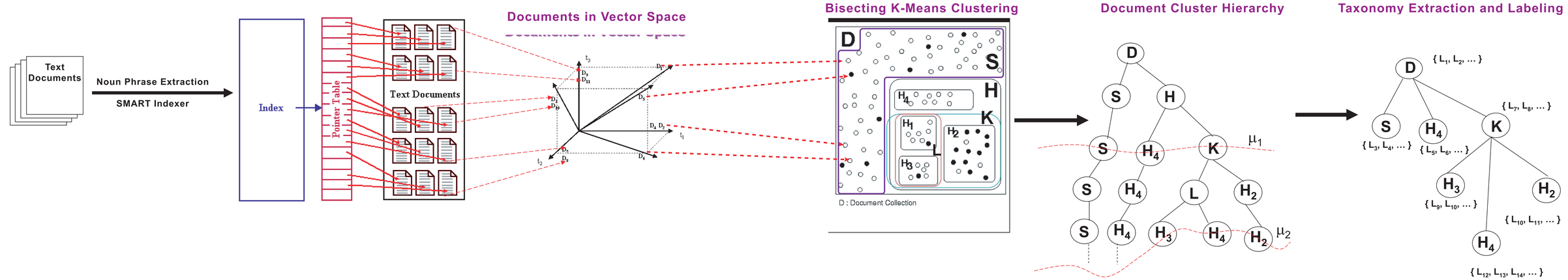
http://cgsb2.nlm.nih.gov/~kashyap/projects/TaxaMiner

Towards (Semi-)automatic Generation of Bio-medical Ontologies

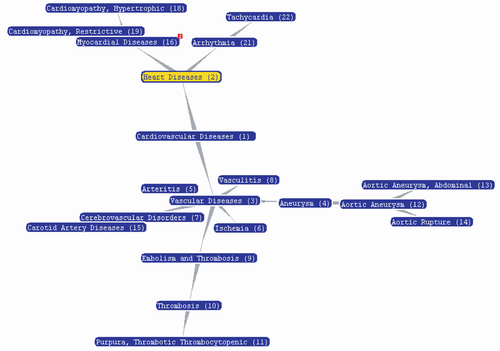
V. Kashyap[1], C. Ramakrishnan[2] and T. C. Rindflesch[1]
[1]National Library of Medicine, Bethesda, MD
[2]LSDIS Lab, Department of Computer Science, UGA, Athens, GA



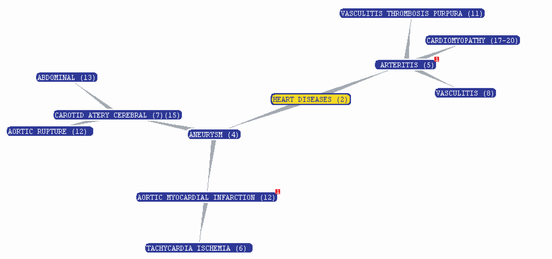
Introduction

- Design and construction of domain taxonomies/ontologies
  - is a very costly activity
- Very human intensive process
- Need to minimize human involvement and other resources
- Combine NLP techniques and Statistical Clustering approaches
- Advantages
  - –Identify gaps in existing biomedical taxonomies such as MeSH
  - –Bootstrap taxonomies and ontologies for new research areas

Gold Standard Taxonomy



Learned Taxonomy



Generated from 3000 documents, clustering upto 20 levels

Discussion

- TaxaMinerProject Goal:
  - –Design, experiment and test algorithms for
    - semi-automatic generation of biomedical ontologies
- Techniques:
  - –Statistical techniques necessary for unsupervised
    - learning
  - –NLP techniques useful for proper labeling of clusters
  - –Initially begin with domain independent techniques
    - and then apply domain specific techniques
- Challenge:
  - –Generalize techniques across multiple domains within
    - and beyond biomedical informatics